复杂场景文本段识别

王孝男，张 利，何思楠

(清华大学 电子工程系，北京 100084)

摘 要： 针对背景复杂或者存在字符黏连时文本段图片无法准确切分的情况进行了研究，提出了一种复杂场景文本段识别方法。该方法利用图像和文字序列的相关性设计双向递归神经网络对图像特征序列进行编码，然后设计集成的连接时间分类（CTC）和注意力（attention）模块对编码特征进行解码输出。该算法在多个数据集（公开数据集 ICDAR2013 和 ICDAR2003 以及验证码数据集）上进行测试，得到识别准确率分别为 90.2%，87.4%和 92.5%，从而证明了该算法的有效性。实验结果对文本段识别和应用有重要意义。

关键词： 文本段识别；连接时间分类；注意力；集成

中图分类号： TP391.41 doi: 10.3969/j.issn.1001-3695.2018.03.0230

Text segmentation based on integration of CTC and attention

Wang Xiaonan, Zhang Li, He Sinan

(*Dept. of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: Text segment recognition was hard due to the complex background and merged characters, to address this problem, this paper proposed a method to recognize text segments in complex scene. Firstly, it designed bidirectional recurrent neural network to encode image feature sequence, based on correlation between images and text sequences. Then it used integrated connected time classification (CTC) and attention(Attention) module as decoder, to decode and output coding features. The method experimented on multiple data sets (public data sets ICDAR2013 and ICDAR2003 and verification code data sets) , the recognition accuracy rates were 90.2%, 87.4% and 92.5%, which demonstrated the method's effectiveness. The experiment results are significant to text segment identification and application.

Key words: text segment recognition; CTC; Attention: integration

0 引言

文本识别指的是将一张包含文本的图片中的文字提取并识别成文字字符串的过程[1]。通常情况下，这个问题可以划分为两个阶段，一个是文本检测阶段，另一个是文本段识别阶段。文本检测将图片中的文本区域检测并用矩形框框出，文本识别则读取文本段图片并识别其内容。

本文关注文本段识别过程，近年来文本段识别取得了一定进展，从单个字符分割识别到实现文本段端到端整体识别[2~4]。字符分割方面，通过阈值寻找投影直方图极值的方法可以实现简单背景下的单字切分；复杂场景下，通过滑窗搜索以及字符间隔分类器判别[5,6]的方法寻找字符分割点也可以实现切分。文本段识别跟单字识别最大的不同还有文字之前具有很强的语义信息，所以要捕捉文字之间的这种语义强相关性。这种情况下，字符切分和单字识别可以结合语言先验信息增强识别效果，结合语言模型可以使用的优化方法如：贝叶斯推断[7~9]，马尔可夫随机场[10，11]，条件随机场[12，13]和概率图模型[14，15]。但是引入语言模型的整体架构并不是端到端的，需要分别训练各个模块，Jaderberg 等人[16]通过将卷积神经网络特征级联递归神经网络实现图像序列化以及序列特征提取，这个架构通过将语言模型这种强语言相关性约束替换成了基于图像序列的递归神经网络，从而实现了端到端训练。Lee 等人[17]在此基础上引入了基于 attention 机制的编解码模型，使得模型可以从图像特征序列中选择特定序列进行解码。Shi 等人[18]引入了语音识别中的 CTC 模块通过动态规划计算真值序列出现的概率，并最大化概率对应的对数损失函数进行训练。Breuel[19]对卷积-递归框架作了改进，对输入加入了几何归一化部分，使得网络整体的识别性能有了进一步提升。Bolan 等人[20]对于卷积模块的输入层加入了梯度投影直方图（HOG）特征，从而在输入通道上加了几个特定的特征图，对于整体性能也有一些效果提升。

收稿日期：2018-03-04；修回日期：2018-04-19

作者简介：王孝男（1993-），男，安徽阜阳人，硕士研究生，主要研究方向为图像处理、模式识别、文字识别（wxn15@mails.tsinghua.edu.cn）；张利（1965-），男，教授，博导，博士，主要研究方向为图像处理.和模式识别；何思楠（1993-），男，江西抚州人，硕士研究生，主要研究方向为图像处理、模式识别和文本检测.

本文通过将 CTC 和 attention 进行集成来识别文本段。Attention 解码机制可以充分利用编码特征序列进行特征筛选，CTC 通过概率计算实现预测值和真实值的对齐。通过使两个网络共用一个编码模块，再对解码层进行损失函数加权累加来实现模块集成。实验结果显示两者集成后的网络可以提高文本段识别的准确率，同时通过注意力权值的可视化过程揭示了文本段识别的内部机制。

1 联合 CTC-Attention 机制

1.1 Attention 编解码模型

编解码模型对于输入特征序列通过底层 RNN 进行特征编码，然后通过上层 RNN 解码网络进行解码输出。基于 attention 注意力机制的编解码模型可以捕捉输入序列中跟输出标签对应的特定部分。有两种注意力机制：硬注意力机制和软注意力机制。硬注意力机制在特定的时间步上通过权值选择特定的一个区域，对应的损失函数存在突变从而导致网络难以训练。软注意力机制在特定的时间步上对所有区域取权值平均，所以更易于进行端到端训练，本文中选用软注意力机制。

分别定义两个 RNN 模块作为编码和解码模块，一个双向 RNN 作为编码模块，对输入的图像特征序列进行编码；另一个双层 RNN 作为解码模块，产生或者解码输出序列。定义编码的隐藏层状态为 $(h_1,h_2,\dots h_{T_A})$，解码层的隐藏状态为 $(d_1,d_2\dots d_{T_B}) \coloneqq (h_{T_A+1},h_{T_A+2}\dots h_{T_A+T_B})$。在每个时间步$t$上根据编码隐藏层状态向量计算注意力向量，定义：

$$u_i^t = v^T tanh(W_1^{'}\ h_i + W_2^{'}\ d_t) \quad (1)$$

$$a_i^t = softmax(u_i^t) \quad (2)$$

$$d_t^{'} = \sum_{i=1}^{T_A} a_i^t h_i \quad (3)$$

向量 v 和权值矩阵 W_1' 和 W_2' 是模型中可训练的参数，v 是一个向量，W_1' 和 W_2' 是权值矩阵。向量 u^t 长度为 T_A，第i 个元素对应注意力集中第i 个编码隐藏状态 h_i 对应的打分值。这些得分值通过 softmax 函数进行正则归一化生成基于编码隐藏状态向量的注意力掩码。最后将 d_t' 和 d^t 级联作为新的隐藏状态输入到下一个时间步的模型中并作为当前时刻的输出预测向量。

对于输入序列 x，真值序列 l，对应的损失函数：

$$loss_{attention} = -lnP(l|x) = -\sum_u \ln P(l_u|x,l_{1:u-1}) \quad (4)$$

其中$l_{1:u-1}$为当前真值标签前的所有字符。

1.2 CTC 序列对齐

CTC 通过引入特定的映射规则，对标签序列进行反映射计算在映射空间中的生成概率从而实现特征序列和标签序列的自动对齐,CTC 忽略了标签序列中每个标签具体对应的位置，而计算标签序列 l 整体在预测结果 $y = y_1,y_2\dots y_T$ 的后验概率。所以当使用该概率值的负对数值作为训练目标时只需要输入图像和对应的标签序列，而避免了标记每个字符对应的位置。CTC 计算条件概率的原理如下：输入标签序列$y = y_1,y_2\dots y_T$，其中 T 对应序列的长度,$y^t \in \Re^{|I'|}$（$I' = I \cup \text{blank}$，I 包含所有的字符标签, blank 为额外的空格标签)。一个序列映射函数 β 定义在 $\pi \in I'^T$，β包含两种操作：首先溢出重复的字符，然后移除空格字符。通过 β 函数将 π 映射为 l。例如 β 映射“—hh-e-l-ll-oo”为“hello”.条件后验概率定义为通过β将所有的 π 映射到 l 的后验概率总和：

$$p(l|y) = \sum_{\pi:\beta(\pi)=l} p(\pi|y) \quad (5)$$

其中:π 出现的后验概率定义为$p(\pi|y) = \prod_{t=1}^{T} y_{\pi_t}^t$，$y_{\pi_t}^t$ 是在时间戳 t 时对应标签 π_t 的概率。直接计算式（5）对应的为时间复杂度较高，通过动态规划方法计算前向-后向向量可以有效将复杂度降低。

对于输入序列 x，真值序列 l，对应的损失函数：

$$loss_{ctc} = -\ln P(l|x) = -\ln P(y|l) \quad (6)$$

1.3 联合 CTC-Attention

联合 CTC-Attention 将 attention 模块和 CTC 模块进行集成，其中 CTC 和 attention 共用一个编码 RNN 模块，网络结构如图 1 所示。这种集成方法，既可以通过 attention 架构中的编解码机制充分利用当前位置以前的所有特征信息，又可以通过 CTC 中通过计算全局概率的方法利用当前位置后的特征信息，从而对特征信息进行充分利用。同时，CTC-attention 通过在 attention 模块中并入 CTC 模块，从而既加速了网络的收敛速度，又提高了网络的识别性能。对 Attention 向量的可视化过程可以揭示网络识别文本段的内部机理。网络的损失函数设计为两个模块损失函数的权值累加，具体公式为

$$loss_{all} = (1-\alpha) * loss_{attention} + \alpha * loss_{ctc} \quad (7)$$

其中 α 为可以学习的参数，范围为$0 \le \alpha \le 1$

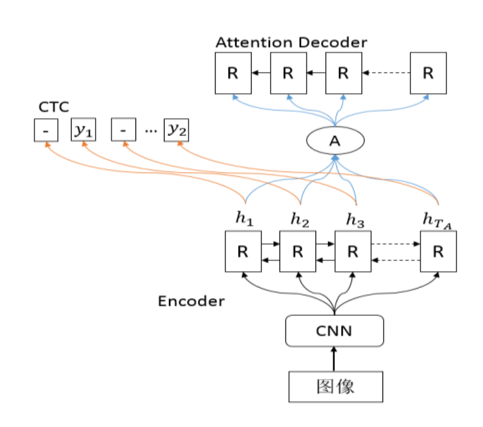


图 1 联合连接时间分类-注意力网络结构图

(R 表示 RNN 单元,A 表示注意力层,CNN 表示卷积神经网络)

2 实验设计

2.1 数据集

本文在三个数据集上进行算法实验。ICDAR2003 数据集，ICDAR2013 数据集，YZM 数据集，使用 Synth90k 数据集作为前两者的训练集。

ICDAR2003（IC03）：街道文本识别数据集，其中测试集包括 251 个全景图片和 860 个切割得到的字段图片；**ICDAR2013（IC13）**：街道文本识别数据集，其中测试集包括 1010 个切割得到的字段图片，大部分图片来自于 ICDAR2003

数据集；**YZM**：验证码数据集，训练集包含两万张验证码图片，测试集包含 1000 张测试验证码图片，图片内容中包含数字和英文字母且背景比较模糊；**Synth90k**：数据集包含 9 百万个合成的切割字段图片。这个合成数据集是高度仿真的文本段图片，可以被用来做为文本段识别的训练集。

2.2 实现细节

实验采用图 1 所示的网络结构设计，CNN 部分采用图 2 所示的架构。编码层为双向的 LSTM，每个 LSTM 单元的隐藏层节点数目为 256；解码层为两个模块的集成，这两个模块分别是：CTC 模块和基于 Attention 的双层 LSTM 解码模块，解码 LSTM 单元的隐藏层节点数目都设置为 128。Attention 向量的计算方法如式（1）所示。

由于输入图片尺寸具有多样性，首先对图片进行归一化。首先保持图片的长宽比不变将图片放缩为高度为 32，宽度为 W。放缩后的图片输入到 CNN 之后输出特征图的长度恰好为 1，可以直接序列化再输入到上层模块中。根据W的大小划分到指定的区间中从而采用对应的解码长度。区间列表为：[64,108]，[108,140]、[140,256]、[256,W_{MAX}](其中 W_{MAX} 表示图片的最大宽度)，对于各个区间的解码长度分别对应为 11,17,19,22,32。在测试过程中采取同样的策略，然后对于W<64 以及W>W_{MAX} 的情况分别进行补零和缩放到宽度为 W_{MAX}。

网络训练采用的方法是随机梯度下降法，通过反向传播来计算每个层的梯度。其中，RNN 部分使用时间反向传播法[21]来计算对应的微分误差。CTC 部分使用前向-后向法[22]来计算对应的微分误差。训练时网络参数初始化使用截尾高斯初始化方法，其中均值为 0.0，截尾标准差为 0.05。网络优化方法使用 Adam 算法，初始学习率为 0.001，beta1 为 0.9，beta2 为 0.999，Adam 算法可以自适应的计算每个时刻的学习率。

评价指标采用的是正确率 $accuracy$，计算方法：

$$accuracy = \frac{M}{N} \quad (8)$$

其中：M 代表数据集中文本段图片识别完全正确的样本数量，N 代表数据集中总样本数量。

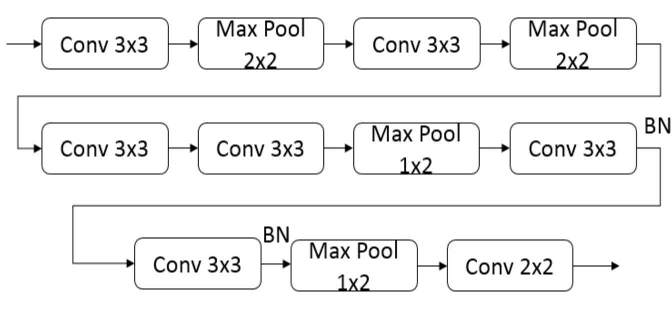


图 2 卷积神经网络具体架构图

2.3 实验结果

分别在 IC03、IC13、YZM 三个数据集上进行实验，前两个数据集包含自然场景文本段图片，最后一个数据集包含人工验证码图片，其特点都是背景复杂，文字前景和背景都比较模糊。对于参数 α 我们分别选择 0.2,0.5,0.7 进行实验对比，识别准确率如表 1 所示。

表 1 文本段数据集实验结果(指标为正确率,单位为%)

算法	IC03	IC13	YZM
CTC	89.4	86.7	90.0
Attention	88.7	87.2	91.3
CTC-Attention(α=0.2)	**90.2**	**87.4**	**92.5**
CTC-Attention(α=0.5)	88.8	86.0	91.0
CTC-Attention(α=0.7)	86.1	85.5	89.4

从表 1 可以看出，CTC 和 attention 在解码端进行集成可以提升文本段识别的整体准确率，其中当 α 为 0.2 时提升幅度最大，此时在三个数据集上均有最好的效果，随着 α 升高整体识别准确率有一定下降趋势。

下面列出本文算法和当前主流算法在两个公开数据集 ICDAR2003 和 ICDAR2013 上的横向对比，结果如表 2 所示。可以看出，本文算法在这两个公开数据集上效果比一些主流算法要好。

表 2 IC03 和 IC13 文本段识别结果(指标为正确率,单位为%)

算法	IC03	IC13
CRNN(CTC)[18]	89.4	86.7
文献[16]	89.6	81.8
PhotoOCR[7]	81.2	82.8
文献[17] (Attention)	88.8	87.2
CTC-Attention(α=0.2)	**90.2**	**87.4**

为了评估算法复杂度和模型的收敛速度，在 Synth90k 数据集上进行训练的过程中每隔一个训练批次输出其在 IC13 测试集上的测试准确率。实验选取参数 α 为 0.2 的模型，同时和 CTC 和 attention 的测试结果作对比，结果如图 3 所示。

从图 3 可以看出，在收敛的情况下，attention 方法比 CTC 方法在测试集上的表现更好，但是 attention 方法存在训练时间复杂度较高、收敛较慢的问题。而 CTC 方法训练过程收敛相对更快，但是其测试准确度不如 attention 方法。CTC-attention 通过在解码 attention 层集成入 CTC 模块，加速了 attention 的训练速度，同时也使整体测试性能得到了提升，其收敛时在测试集上的表现比这两个方法都要好。

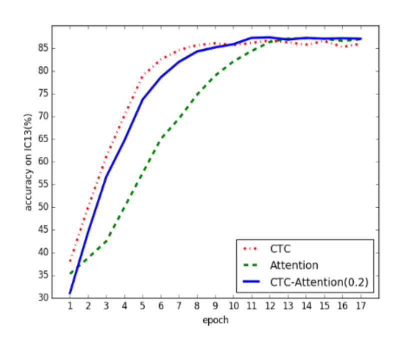


图 3 模型训练曲线对比图

(横轴表示训练批次,纵轴表示 IC3 测试准确率)

2.4 可视化分析

为了更好地揭示网络的工作原理，通过对 attention 解码层的权值可视化来分析网络的工作机制。对于式（2）中的 a_i^t 表示

编码特征序列中第i个编码向量作用于第t个解码向量的权值大小。们通过对每个有效时间步进行 a^t 的大小可视化可以分析出对于当前解码时间步哪些区域其起着主要作用。

基于上述分析，图 2 中对输入的图片进行 CNN 特征提取和双向 LSTM 编码后得到的特征序列，经过 attention 层后对编码特征序列进行权值累加导入到解码 LSTM 中。通过权值向量进行可视化可以看出上述过程中对于特定的解码位置哪些特征区域起着主要作用。图 4 中当对应区域的编码特征序列起作用的程度越大，图片中对应的像素值越高，相应的区域就会越亮。可视化结果揭示了文本段识别的内部机理，对于识别结果中的每个字符，恰好是其相应位置的图像区域起着主要作用，Attention 通过对不同区域分配不同的权值来体现对特定标签字符的“注意力”。

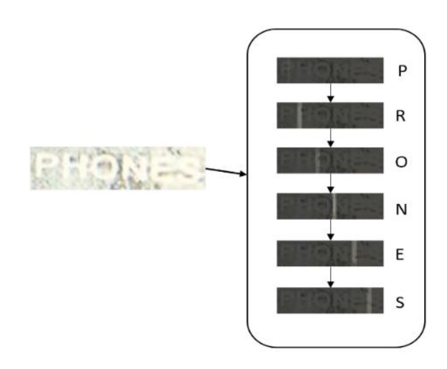


图 4　注意力权值可视化

3　结束语

文本段识别的传统识别方法是先切分再结合语言模型进行后处理，这个过程会造成每一步误差的累积，影响整体识别的准确率。文本分析了设计端到端网络进行文本段识别的必要性，对于字符之间存在粘连或者文字图片的背景比较复杂时，传统字符切分方法的错误率较高，采用整体端到端识别可以较好的解决这种难切割问题。文本分析了目前两种端到端的方法（CTC 和 attention）的原理以及各自的优缺点，设计了一个集成网络进行文本段端到端识别。在多个数据集上的实验结果表明，通过采用共享编码层以及不同解码层的集成，该网络可以提高文本段识别的整体准确率。本文对网络进行文本段图片识别的工作机制进行了可视化分析，对网络解码过程进行直观的理解和认识。

参考文献：

[1] 王科俊. 中文印刷体文档识别技术 [M]. 北京: 科学出版社, 2010. (Wang Kejun. Chinese printed document recognition technology [M]. Beijing: Science Press, 2010.)

[2] 吴锐. 自然场景中文本识别技术研究及实现 [D]. 哈尔滨: 哈尔滨工业大学, 2010. (Wu Rui. Research and implementation of text recognition technology in natural scenes [D]. Harbin: Harbin Institute of Technology, 2010.)

[3] 张当中. 汉字识别技术综述 [J]. 语言文字应用, 1997 (2): 79-88. (Zhang Dangzhong. An overview of chinese character recognition technology [J]. Linguistic Writing, 1997 (2): 79-88.)

[4] Ye Q, Doermann D. Text detection and recognition in imagery: a survey [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2015, 37 (7): 1480-1500.

[5] 潘炜深, 金连文, 冯子勇. 基于多尺度梯度及深度神经网络的汉字识别 [J]. 北京航空航天大学学报, 2015, 41 (4): 751-756. (Pan Weishen, Jin Lianwen, Feng Ziyong. Chinese character recognition based on multi-scale gradient and deep neural network [J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41 (4): 751-756.)

[6] 贺欣. 自然场景文字切分和文本行识别方法研究 [D]. 北京: 中国科学院大学, 2016. (He Xin. Research on text segmentation and text line recognition in natural scenes [D]. Beijing: University of Chinese Academy of Sciences, 2016.)

[7] Bissacco A, Cummins M, Netzer Y, *et al.* PhotoOCR: reading text in uncontrolled conditions [C]// Proc of IEEE International Conference on Computer Vision. 2013: 785-792.

[8] Zhang Dongqing, Chang Shihfu. A Bayesian framework for fusing multiple word knowledge models in videotext recognition [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2003: 528-533.

[9] Weinman J, Learned-Miller E. Improving recognition of novel input with similarity [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern RecognitionIEEE Computer Society Conference on. 2006: 308-315.

[10] Chen Datong and Odobez J. Video text recognition using sequential monte carlo and error voting methods [J]. Pattern Recognition Letter, 2005, 26 (9): 1386-1403.

[11] Weinman J, Learned-Miller E, Hanson A. A discriminative semi-Markov model for robust scene text recognition [C]// Proc of International Conference on Pattern Recognition. 2008: 1-5.

[12] Jawahar C V. Top-down and bottom-up cues for scene text recognition [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2012: 2687-2694.

[13] Shi Cunzhao, Wang Chunheng, Xiao Baihua, *et al.* Scene text recognition using part-based tree-structured character detection [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2013: 2961-2968.

[14] Wachenfeld S, Klein H U, Jiang Xiaoyi. Recognition of screen-rendered text [C]// Proc of International Conference on Pattern Recognition. 2006: 1086-1089.

[15] Lee S H, Kima J H. Complementary combination of holistic and component analysis for recognition of low-resolution video character images [J]. Pattern Recognition Letters, 2008, 29 (4): 383-391.

[16] Jaderberg M, Simonyan K, Vedaldi A, *et al.* Deep structured output learning

for unconstrained text recognition [J]. Eprint Arxiv, 2015, 24 (6): 603–611.

[17] Lee C Y, Osindero S. Recursive recurrent nets with attention modeling for ocr in the wild [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2231-2239.

[18] Shi Baoguang, Bai Xiang, Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39 (11): 2298-2304.

[19] Breuel T M. High performance text recognition using a hybrid convolutional-LSTM implementation [C]// Proc of IAPR International Conference on Document Analysis and Recognition. Washington DC: IEEE Computer Society, 2017: 11-16.

[20] Su Bolan, Lu Shijian. Accurate recognition of words in scenes without character segmentation using recurrent neural network [J]. Pattern Recognition, 2016, 63: 397-405.

[21] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Netw, 2005, 18 (5): 602-610.

[22] Graves A, Gomez F. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]// Proc of International Conference on Machine Learning. 2006: 369-376.